Systematic review of measurement properties of questionnaires measuring somatization in primary care patients

Article (Accepted Version)

SYSTEMATIC REVIEW OF MEASUREMENT PROPERTIES OF QUESTIONNAIRES MEASURING SOMATIZATION IN PRIMARY CARE PATIENTS

Kate Sitnikova, MSc [1*], Sandra MA Dijkstra-Kersten, MSc [1], Lidwine B Mokkink, PhD [2], Berend Terluin, MD, PhD [1], Harm WJ van Marwijk, MD, PhD [3], Stephanie S Leone, PhD [4], Henriëtte E van der Horst, MD, PhD [1], Johannes C van der Wouden, PhD[1]

Email:
Kate Sitnikova: e.sitnikova@vumc.nl
Sandra MA Dijkstra-Kersten: s.kersten@vumc.nl
Lidwine B Mokkink: w.mokkink@vumc.nl
Berend Terluin: b.terluin@vumc.nl
Harm WJ van Marwijk: harm.vanmarwijk@manchester.ac.uk
Stephanie S Leone: sleone@trimbos.nl
Henriëtte E van der Horst: he.vanderhorst@vumc.nl
Johannes C van der Wouden: j.vanderwouden@vumc.nl

Address:
[1] Department of General Practice and Elderly Care Medicine, Amsterdam Public Health Research Institute, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, the Netherlands
[2] Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, the Netherlands
[3] Center for Primary Care, Institute of Population Health, University of Manchester, United Kingdom
[4] Department of Public Mental Health, Trimbos Institute: Netherlands Institute of Mental Health and Addiction, Da Costakade 45, 3521 VS Utrecht, the Netherlands

* Corresponding author

Abstract

Objective: The aim of this review is to critically appraise the evidence on the measurement properties of self-report questionnaires measuring somatization in adult primary care patients and to give recommendations about which questionnaires are most useful for this purpose.

Methods: We assessed the methodological quality of included studies using the Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist. To draw overall conclusions about the quality of the measurement instruments, we conducted an evidence synthesis using predefined criteria for good measurement properties.

Results: We found 24 papers, investigating 9 measurement instruments. Studies on the Patient Health Questionnaire-15 (PHQ-15) and the Four-Dimensional Symptom Questionnaire (4DSQ) somatization subscale were found most frequently and covered the broadest range of measurement properties. These questionnaires had the best internal consistency, structural validity, construct validity and test-retest reliability. The PHQ-15 also had good criterion validity, whereas the 4DSQ somatization subscale was validated in several languages. The Bodily Distress Syndrome (BDS) checklist had good internal consistency and structural validity. Some evidence was found for good construct validity and criterion validity of the Physical Symptom Checklist (PSC-51) and good hypotheses testing of the Symptom Check-List (SCL-90-R) somatization subscale. However, the three latter questionnaires were only studied in a small number of studies each.

Conclusion: Based on our findings, we recommend the use of either the PHQ-15 or 4DSQ somatization subscale as an outcome measure for somatization in primary care. Other questionnaires, such as the BDS checklist, PSC-51 and the SCL-90-R somatization subscale show promising results but have not yet been studied extensively in primary care.

Keywords: Measurement properties, Primary care, Self-report questionnaire, Somatization

Introduction

Somatization is defined as "a tendency to experience and communicate somatic distress and symptoms unaccounted for by pathological findings, to attribute them to physical illness, and to seek medical help for them" (1). Multiple explanatory models have been proposed for this phenomenon, such as somatosensory amplification, sensitization, endocrine dysregulation, dysfunctional disease beliefs, avoidance of physical, mental and social activity, and abnormal proprioception (2). Although experiencing one or several medically unexplained symptoms without a known underlying medical explanation is common for all people, especially in stressful situations, experiencing many medically unexplained symptoms from various organ systems implies somatization (3). If symptoms persist, patients may seek medical help. This is common for people in all health care settings, but particularly so in primary care (4, 5). Due to its generalist nature, primary care is the first port-of-call for people who are worried about such physical experiences.

Physical symptoms in primary care can be aligned across a spectrum of the number, severity and functional impairment of symptoms, with having just one or a few transient symptoms at one end of the spectrum, and having multiple severe symptoms for a long period of time and therefore meeting diagnostic criteria for a somatoform disorder according to the Diagnostic and Statistical Manual of mental disorders 4th, (DSM-IV) (6) or 5th edition (DSM-5) (7), at the other end (8).

Experiencing multiple physical symptoms is an imperative part of somatization, therefore we restrict our definition of 'somatization' to having multiple physical symptoms at the same time and look into questionnaires that quantify these symptoms as a proxy for somatization. We acknowledge the various possible explanatory factors and consequences that somatization can have, but do not focus on these in the current study.

Discussing the presence and number of the most common physical symptoms with self-report questionnaires can be a valuable tool for psycho-education and shared decision-making, particularly in primary care, as the clinical dialogue could generate different treatment modalities, such as cognitive behavioural therapy (9), than would have been provided for a condition with a known biological cause. The sooner high levels of somatization are signalled and discussed, the sooner patients can learn to make sense of them and the sooner appropriate care can be provided. As a result, otherwise potentially unnecessary, costly, medical procedures with possible side-effects can be avoided. Considering the general practitioners' (GP) and nurse practitioners' time-restrictions, self-report questionnaires can be a useful, quick, non-invasive tool to assist GPs in detecting symptoms of somatization directly from the patient's point of view. Research comparing the quality of various available measurement instruments to measure somatization in primary care has not yet been done. Therefore, to date, it remains unclear which measurement instrument can be used best for this purpose.

Two previous studies (10, 11) provided overviews of measurement instruments, one for common somatic symptoms (10) and the other for somatoform disorders (11). However, neither was specifically focussed on use in primary care and neither used the state-of-the-art COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) methodology (12, 13) for conducting systematic reviews on measurement instruments.

The aim of this review is to critically appraise the evidence on the measurement properties of (subscales of) self-report questionnaires measuring somatization as an outcome measure in adult primary care patients and to give recommendations about which questionnaires are most useful for this purpose.

Methods

This review is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (14).

Literature search

A search was performed on August 13, 2015 in PubMed/Medline, Embase, Psycinfo and Cinahl from inception. In all databases search terms for construct, population, measurement properties and setting were combined using the Boolean term 'AND'. In PubMed a validated search filter was used for finding articles investigating measurement properties (15). In the other databases, adapted versions of this search filter were used. The adaptations were performed by a scientific information specialist. The full search strategies for each database can be found in Appendix A. A second updated search was performed on October 31, 2016 following the same procedure. Reference lists of the included articles and reviews found during the searches, were searched to identify additional relevant articles. Authors of studies were contacted in case manuscripts of the studies were not available online.

Inclusion and exclusion criteria

Inclusion criteria were:

1. The questionnaire or subscale should aim to measure somatization defined as having multiple physical symptoms.
2. The study population should be adults (age 18 and above) who are patients in primary care.
3. The instrument of study should be developed as a paper or online self-report questionnaire.
4. The aim of the study should be the development of a measurement instrument or the evaluation of one or more of its measurement properties.
5. The study should be published as a full text original article.

Exclusion criteria were:

1. Studies in languages other than English or Dutch.
2. Studies measuring somatization as a trait, rather than a state.
3. Studies investigating a specific functional syndrome (e.g. fibromyalgia, irritable bowel syndrome, chronic pain syndrome).
4. Questionnaires including items on somatization among other items, but without a separate subscore for somatization.

Selection procedure

The selection of articles based on titles and abstracts was independently performed by two reviewers (KS and SDK). Afterwards, these two reviewers separately checked whether the full text articles met

the inclusion criteria. In case of disagreement or doubt, a third reviewer (JW/BT) was consulted in order to make the decision regarding inclusion of the article.

Data extraction

Two reviewers (KS and SDK) independently extracted and evaluated the general characteristics of the measurement instruments, the characteristics of the studies, and information on generalizability and interpretability, using a structured form. When not enough information could be obtained from the included articles, original development articles were consulted. Disagreement between reviewers was discussed until consensus was reached.

Assessment of the methodological quality of the included studies

The methodological quality of the studies was assessed using the COSMIN checklist (12). The COSMIN checklist was developed in an international Delphi Study and can be used to evaluate the methodological quality of studies on measurement properties. The COSMIN checklist consists of 12 dimensions. Nine dimensions contain standards for quality of the methodological properties reliability, measurement error, content validity, structural validity, hypotheses testing, cross-cultural validity, criterion validity and responsiveness. One dimension contains standards for studies on interpretability. One dimension contains general requirements for articles using item response theory (IRT), and one dimension contains general requirements for the generalizability of results.

We used the COSMIN checklist (13) to determine which measurement properties were evaluated in a study. Two reviewers (KS and SK) then independently evaluated the quality of the included studies per measurement property, using the COSMIN checklist 4-point rating scale (12, 16) (available from the website www.cosmin.nl). In case of disagreement or doubt, a third reviewer (JW/LM) was consulted

We modified the COSMIN checklist slightly by omitting the two items on the percentage and handling of missing data. This choice was made because it is unclear how missing data contributes to methodological quality and what the best way to handle missing data is, when investigating measurement properties. It is also possible that there are no missing data at all. Also, when no information on missing data is given, it does not necessarily mean that missing items were not handled well and a lower rating of methodological quality can be given unjustifiably.

Furthermore, the cross-cultural validity dimension concerns two different aspects: 1) translation of the instrument, and 2) the actual cross-cultural validation analysis between two culturally different groups. To acknowledge these two aspects, we decided to split the dimension into two sections, i.e. translation score (items 4-11) and the cross-cultural validation score (items 1-3 and 12-15). For criterion validity, we considered validated interviews based on the Diagnostic and Statistical Manual of mental disorders (DSM-IV) (6) criteria for somatoform disorders to be the gold standard (e.g. the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (17), the Structured Clinical Interview for DSM-IV-TR Axis I Disorders (SCID-I) (18), and the Mini International Neuropsychiatric Interview (MINI) (19)). Studies using other comparison instruments were not considered to address criterion validity but evaluated under hypotheses testing.

Evaluation of the study results against criteria for good measurement properties

The results of each study were extracted and compared to criteria for good measurement properties developed by Prinsen and colleagues in cooperation with the COSMIN initiative (20). We slightly adjusted the criteria described by Prinsen et al. for reliability and criterion validity (Appendix B). The adjustment was done because various studies reported other appropriate coefficients besides the intraclass correlation coefficient (ICC) or weighted kappa when assessing reliability, and other appropriate values than a correlation were used when assessing criterion validity. For reliability we scored a '+' when ICC or weighted kappa was ≥ 0.70 and also when Pearson's correlation coefficient was ≥0.80. We scored a "-" when these criteria were not met. For criterion validity we scored a '+' when the correlation with the gold standard was ≥0.70, but also when the area under the curve (AUC) was ≥ 0.70 or, in case no correlation or AUC was provided, when both sensitivity and specificity were ≥60%. We scored a "-" when none of these criteria were met.

Data syntheses

For each measurement instrument, the overall levels of evidence on each measurement property were synthesized using the results on measurement properties from all included studies (21). The levels of evidence were adjusted for the methodological quality of each study. The levels of evidence used are provided in Table 1.

Table 1. Levels of evidence for the quality of the measurement properties (21)

Level	Rating	Criteria
Strong	+++ or ---	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
Moderate	++ or --	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
Limited	+ or -	One study of fair methodological quality
Conflicting	+/-	Conflicting findings
Unknown	?	Only studies of poor methodological quality

+ = positive rating, ? = indeterminate rating, - = negative rating

Results

Included studies

The search yielded 5318 hits in total, of which 1326 hits in PubMed, 3029 in Embase, 729 in Psycinfo and 234 in Cinahl. An overview of the searches and selection of articles is presented in Figure 1.

After removing duplicates, a total of 4129 articles remained. After screening titles and abstracts, 151 articles remained and were assessed for eligibility based on their full-texts. Twenty-one articles were eligible for inclusion. After screening the reference list of these 21 articles, and of several reviews

found in the search, three more eligible articles were identified. This resulted in a total of 24 eligible articles describing 52 studies on 9 different measurement instruments. When using the term 'study', we refer to the investigation of one single measurement property at a time. Various studies on different measurement properties may be described in a single article.

Internal consistency was assessed in 15 studies, reliability was assessed in 3, measurement error in 1, structural validity in 9, hypotheses testing in 13, cross-cultural validity in 3, criterion validity in 7 and responsiveness in 1. Content validity was not assessed in any of the studies.

The characteristics of the studies are provided in table 2.

Description of the questionnaires

Table 3 summarizes the general characteristics of the included questionnaires. Twelve studies assessed the somatic symptom scale of the Patient Health Questionnaire (PHQ), i.e. the Patient Health Questionnaire-15 (PHQ-15) (22-28), and one study assessed the brief PHQ-r (29), which is a Turkish version of the PHQ designed for the particular study included in this review (29). The brief PHQ-r consists of four subscales: somatoform disorder, depressive disorders, panic disorder and functioning of the patient. In our review we only looked at the somatoform disorder subscale. Twenty-two studies described in eight papers assessed the measurement properties of the Four-Dimensional Symptom Questionnaire (4DSQ) somatization subscale (3, 30-36). The entire 4DSQ consists of 4 subscales: distress, depression, anxiety and somatization. As explained above, we restricted ourselves to the psychometric properties of the subscale somatization. Four studies assessed the Symptom Check-List-90-R (SCL-90-R) somatization subscale (37, 38). The SCL-90-R is a comprehensive questionnaire that aims to measure a broad range of psychological problems. It consists of 9 subscales: somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, anger-hostility, phobic anxiety, paranoid ideation and psychoticism. Again, we only assessed the subscale somatization. The other measurement instruments, i.e. the Schedule for Evaluating Persistent Symptoms (SEPS) (39), the Physical Symptom Checklist (PSC-51) (40), the Common Mental Disorders Questionnaire (CMDQ) (41), the Ghent Multidimensional Somatic Complaints Scale (GMSCS) (42), the Screening for Somatoform Symtptoms-2 (SOMS-2) (43) and the Bodily Distress Syndrome (BDS) checklist (44) were assessed in one to three studies each. The SEPS records medically unexplained symptoms. The PSC-51 measures somatoform disorders and is based on the DSM-III classification (45), which is an outdated version of the DSM criteria. The CMDQ is a diagnostic tool for common mental disorders and it consists of three subscales: somatoform disorder, mental disorder and alcohol dependence. In this review we will only look into the somatoform disorder subscale. The items for the somatoform disorder part are taken from the Symptom Checklist (SCL-90) (46)and the Whiteley index (47). The GMSCS assesses somatic complaints. The SOMS-2 , is originally developed in German (48, 49), but in the study included in this review, the authors investigated an adapted Portuguese version of the SOMS-2 and a shorter version, the R-SOMS-2. Finally, the BDS aims to diagnose functional disorders. Three papers on the 4DSQ were written in Dutch (34-36). All other articles were in English. The measurement properties of the PHQ-15, 4DSQ somatization subscale and SCL-90-R somatization subscale were investigated by seven, five and two different research teams, respectively. The measurement properties of the remaining questionnaires were investigated by a single research team each.

Methodological quality of the included studies on measurement properties

The methodological quality of studies investigating internal consistency and criterion validity varied widely from poor to excellent. All studies investigating structural validity and cross-cultural validity were of excellent quality. Most studies investigating hypotheses testing, on the other hand, were of fair quality, because hypotheses were frequently not stated explicitly.

Measurement properties of the questionnaires and evidence rating

Internal consistency

Table 4 provides an overview of the 15 studies that assessed internal consistency (3, 23, 24, 26, 31, 32, 34-39, 42-44). All questionnaires except the SCL-90-R somatization subscale, SEPS and the (R)-SOMS-2 showed good internal consistency, which was supported by strong evidence. Results for the SEPS and the (R)-SOMS-2 could not be evaluated due to poor quality of the studies.

Test-retest reliability

Table 5 provides an overview of the 3 studies in which the test-retest reliability was assessed (26, 35, 43). The PHQ-15 and the 4DSQ somatization subscale were the most reliable questionnaires, although the study investigating test-retest reliability of the PHQ-15 was of good methodological quality, while the study investigating reliability of the 4DSQ was of fair quality. The evidence on the (R)-SOMS-2 could not be interpreted due to poor quality of the study.

Measurement error

Only one study of poor quality, on the 4DSQ somatization subscale, investigated measurement error (3). The study sample consisted of 1424 participants and showed the following results: standard error of measurement (SEM)=2.80, smallest detectable change (SDC)=7.76. However, the SEM was estimated based on the Cronbach's alpha, which is a method of poor quality according to the COSMIN checklist (12), and consequently, no conclusions about the measurement error could be drawn.

Structural validity

Table 6 summarizes the results for structural validity from the 9 included studies (3, 25, 31, 33, 34, 36, 39, 42, 44). Strong evidence was found for good structural validity of the PHQ-15, the 4DSQ somatization subscale and BDS checklist and for poor structural validity of the GMSCS and the SEPS.

Hypotheses testing

The results from the 13 studies evaluating hypotheses testing (3, 22-24, 30, 34-40, 43) are provided in table 7. Good construct validity was supported by moderate evidence for the PHQ-15, the 4DSQ somatization subscale and the SCL-90-R somatization subscale. Limited evidence supported good construct validity of the SEPS and the PSC-51. The (R)-SOMS-2 seemed to have poor construct validity due to low sensitivity, however, due to poor quality of the study, the evidence could not be interpreted.

Cross-cultural validity

Cross-cultural validity was assessed in three studies on the 4DSQ somatization subscale (31-33). Translation of the questionnaires was not described in these studies. The studies, however, validated

the questionnaire in an English (31), Polish (32) and French (33) population. Strong evidence supported a good validation score. In all studies, results showed that the translated versions conveyed the same meaning as the original Dutch version of the questionnaire, i.e. none of the items included in any of these questionnaires showed differential item functioning between language groups. Also, the same cut-off points for determining severity of somatization could be used across language groups.

Criterion validity

Table 8 summarizes results from the 7 studies investigating criterion validity (22, 26-29, 40, 41). Strong evidence was found for good criterion validity of the PHQ(-15) and limited evidence for good criterion validity of the PSC-51. The CMDQ somatoform disorder subscale seems to have poor criterion validity, although the evidence was limited due to fair methodological quality of the study.

Responsiveness

Responsiveness was evaluated in one study (3) of fair quality on the 4DSQ somatization subscale. In this study 86 GP patients with psychosocial problems (age 40.2 (10.0) and 66% female) completed the 4DSQ twice within a, relatively short, time interval of 10 days. Fifty-nine of these patients answered a 5-point Global Impression (GI) question. The correlation between the somatization change scores and the GI score was weak (r=0.30 (0.04 – 0.53)). The patients were then divided into 2 groups: improved and not improved, and receiver operating characteristic (ROC) analyses were performed. The area under the curve (AUC) was 0.69, just below the cut-off of 0.70 for good responsiveness (50). This limited evidence therefore demonstrated poor responsiveness.

Discussion

Summary of evidence

We identified 24 articles describing 52 studies on measurement properties of 9 self-report questionnaires measuring somatization in the primary care setting. The PHQ-15 and the 4DSQ somatization subscale were studied the most, in 13 and 22 studies respectively. The SCL-90-R somatization subscale was described in four studies and the remaining questionnaires were described in only one to three studies per measurement instrument, which weakens the level of evidence with which the results were interpreted.

Based on our overview, we recommend using the PHQ-15 or 4DSQ as an outcome measure of somatization. The choice between the PHQ-15 and 4DSQ somatization subscale seems somewhat equal and can be based on practical considerations. The two measurement instruments have nearly the same number of items, however the PHQ-15 enquires about symptoms in the previous four weeks, whereas the 4DSQ has a recall period of one week. Having to recall symptoms over a longer period of time could cause more recall bias. However, reporting symptoms from the previous week only, could possibly leave out important information about symptoms that were present previously, but by chance were less prominent in the past week. Also, the PHQ-15 includes two items enquiring about symptoms linked to menstruation and sexual intercourse. The 4DSQ does not include items on these topics. The choice for one of the two instruments could therefore depend on the patient population. For instance, when using a questionnaire with female patients within their reproductive

age range, the PHQ-15 could provide useful information. With older patients, for instance women after having reached menopause, the 4DSQ could possibly be more suitable. A health care provider interested in screening for the DSM-IV somatoform disorder may opt for the PHQ-15, as this instrument has been compared against this diagnosis. On the other hand, the 4DSQ somatization subscale may be more suitable for Polish, French and Dutch speaking patients due to its validation within these population groups.

A promising measurement instrument is the BDS checklist. Although its measurement properties were only investigated in two studies, strong evidence was found for good internal consistency and structural validity. However, more research is needed to investigate the quality of the remaining measurement properties.

Based on the studies included in this review the SEPS, GMSCS and the CMDQ somatoform disorder subscale seem less suitable for measuring somatization in primary care, due to poor structural validity of the first two instruments, and poor criterion validity of the latter. However, more research of these measurement instruments is needed to be able to draw more solid conclusions about their quality.

The remaining measurement instruments, the SCL-90-R somatization subscale, PSC-51 and the (R)-SOMS-2 have been described in a small number of studies where only two measurement properties were assessed. The studies on the (R)-SOMS-2 were of poor quality, therefore, no conclusions were drawn about it in this review. Limited to moderate evidence supported findings of several good measurement properties of the SCL-90-R somatization scale and the PSC-51. Studies on the SCL-90-R somatization subscale in other populations (general population and various secondary care patients) show acceptable to good psychometric properties (51-53). So once again, more research on these measurement instruments in primary care is needed. A point of consideration is that the PSC-51 and the (R)-SOMS-2 consist of 51 and 46 (or 29 in the short version) items, respectively, and are therefore time-consuming .

.

Embedding in existing literature

One previous review provided an overview of diagnostic measurement instruments for somatoform disorders (11). However, this review only focused on the assessment of somatoform disorders, which are at the most severe spectrum of symptoms of somatization and therefore their findings only cover part of the broad range of symptom severity that is encountered in primary care. Also, although measurement properties of the measurement instruments were mentioned, no structured, thorough evaluation of measurement properties was made.

Another previous study provided an overview of self-report questionnaires for common somatic symptoms for use in large-scale epidemiological studies in any type of population, so not specifically for primary care (10). The authors recommend the use of the PHQ-15SCL-90 somatization subscale. However, their aim was to determine which questionnaire was most suitable for research purposes in large-scale studies, and not in health care practice settings. Other than investigating aspects of reliability and validity, the authors focussed on applicability in large-scale studies by examining low burden to participants, to investigators with no specific expertise in the assessment of somatization,

and relevance for the near future. In addition, two important measurement properties, i.e. measurement error and responsiveness were not taken into account in their review. In our review we also found positive results for the PHQ-15 in primary care, but less so for the SCL-90-R somatization subscale. Instead we recommend the 4DSQ somatization subscale, along with the PHQ-15.

Strengths and limitations

The most important strength of this study is that we used the COSMIN taxonomy for deciding which measurement properties were assessed, and that we took the methodological quality of each individual measurement property into account when interpreting the results of the studies, and drawing conclusions on the quality of the included measurement. This provided a structured instrument for assessing all questionnaires in a consistent way.

We modified the COSMIN standard somewhat, by omitting the missing data items from the 4-point rating scale. This modification had consequences for the evidence rating of the questionnaires. Although all results remained in the same direction, the results were of stronger quality due to the modification and stronger conclusions were drawn because of this. This was especially the case for the measurement properties 'internal consistency', 'structural validity' and 'criterion validity'.

A point of consideration is our definition of the term 'somatization', in which we only quantified the multiple physical symptoms. The widely used definition by Lipowski (1) also incorporates cognitions and behaviour of the patients with regard to their multiple physical symptoms. However, as virtually no questionnaire enquires about all those aspects simultaneously, it was impossible to include questionnaires measuring somatization according to Lipowski's definition. Therefore, we chose to focus on the measurement of experienced multiple physical symptoms which can function as a proxy for somatization. However, to cover the entire original definition of somatization, future studies could take the psychological and behavioural aspects of somatization into account as well.

Another point of consideration is that unfortunately there is no true gold standard for somatization. A relatively objective measure that approaches a gold standard is a diagnosis of a somatoform disorder according to the DSM-IV criteria. We therefore referred to these criteria as the gold standard in this review as well. However, it must be noted that these diagnoses only cover the extreme end of the spectrum of somatization. As there is in fact no gold standard for somatization, the comparison with the DSM-IV criteria could be considered hypotheses testing rather than criterion validity.

A third limitation is that we did not include grey literature such as dissertations and conference abstracts. This choice may have contributed to selection bias. Also, due to practical reasons we did not use indirect evidence from studies in which the measurement instruments were actually used. Finally, we excluded full-text articles that were written in a language other than English or Dutch.

A final point of consideration is that we chose to limit our search to studies on questionnaires that were developed or studied in primary care. Questionnaires studied in other populations, such as the general population, community samples, students, secondary care, have therefore been excluded, as measurement properties of measurement instruments may be different in different populations and settings. However, it is possible that some questionnaires could be useful, but have not yet been

studied in primary care, such as the Somatic Symptom Index (SSI) (54),the SOMS-7 (55) or the Somatic Symptom Scale-8 (SSS-8) (56). The latter questionnaire for instance performs similarly to the PHQ-15 in secondary care and has less items. Also, more studies are available on the SCL-90-R somatization subscale in the general population, secondary care and psychiatric populations (51-53, 57). Validation of these questionnaires in primary care may yield interesting information.

Conclusions

The PHQ-15 and the 4DSQ somatization subscale have been studied most in primary care and show the most positive results on a broad array of measurement properties. We therefore recommend the use of one of these two instruments for measuring somatization in primary care. The choice of a preferred measurement instrument can differ depending on the measurement properties that have priority to the user of the questionnaire. Health care providers interested in the closest approximation to a somatoform disorder may favour the PHQ-15, whereas health care providers seeking the best questionnaire for Polish, French or Dutch-speaking patients may choose the 4DSQ somatization subscale instead. Other questionnaires, such as the BDS checklist, SCL-90-R somatization subscale and PSC-51 could benefit from further study primary care. However, the BDS checklist and the PSC-51 consist of a larger number of items than the PHQ-15 and 4DSQ somatization subscale and are therefore more time-consuming. Finally, measurement properties such as measurement error, content validity, cross-cultural validity and responsiveness should be studied further in all measurement instruments using sound research methods.

Author contributions

All authors read and commented on draft versions of the manuscript and approved the final version.

Funding

This project is funded by ZonMw (dossier number 80-83700-98-42070), the Netherlands Organisation for Health Research and Development. The funding body did not and will not have any role in the collection, analysis, or interpretation of data, nor in the writing of the manuscript.

Conflict of interest

Lidwine Mokkink is one of the developers of the COSMIN checklist. Berend Terluin is the developer of the 4DSQ. He did not take any part in assessing the methodological quality of the measurements instruments. The other authors declare that they have no conflict of interest.

Acknowledgments

The authors thank Caroline Terwee for advising on the search strategy. We also thank scientific information specialist René Otten for translating the validated search filter for PubMed in search filters for other databases.

References

1. Lipowski ZJ. Somatization: The concept and its clinical application. Am J Psychiatry. 1988;145:1358-68.
2. van Ravenzwaaij J, olde Hartman TC, van Ravesteijn H, Eveleigh R, van Rijswijk E, Lucassen PLBJ. Explanatory models of medically unexplained symptoms: a qualitative analysis of the literature. Ment Health Fam Med. 2010;7:223-31.
3. Terluin B, van Marwijk HW, Ader HJ, de Vet HC, Penninx BW, Hermens ML, et al. The Four-Dimensional Symptom Questionnaire (4DSQ): a validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. BMC Psychiatry. 2006;6:34.
4. Arnold IA, de Waal MW, Eekhof JA, van Hemert AM. Somatoform disorder in primary care: course and the need for cognitive-behavioral treatment. Psychosomatics. 2006;47(6):498-503.
5. de Waal MWM, Arnold IA, Eekhof JAH, Van Hemert AM. Somatoform disorders in general practice: prevalence, functional impairment and comorbidity with anxiety and depressive disorders. Br J Psychiatry. 2004;184(6):470-6.
6. APA. Diagnostic and statistical manual of mental disorders, 4th ed. (DSM-IV). Washington, DC: American Psychiatric Association; 1994.
7. Diagnostic and Statistical Manual of Mental Disorders. Arlington, VA: American Psychiatric Publishing; 2013.
8. Jackson JL, Kroenke K. Prevalence, impact, and prognosis of multisomatoform disorder in primary care: a 5-year follow-up study. Psychosom Med. 2008;70(4):430-4.
9. van Dessel N, den Boeft M, van der Wouden JC, Kleinstäuber M, Leone SS, Terluin B, et al. Non-pharmacological interventions for somatoformdisorders and medically unexplained physical symptoms (MUPS) in adults. Cochrane Database Syst Rev. 2014;11, CD01114.
10. Zijlema WL, Stolk RP, Lowe B, Rief W, BioShaRe, White PD, et al. How to assess common somatic symptoms in large-scale studies: a systematic review of questionnaires. J Psychosom Res. 2013;74(6):459-68.
11. Hiller W, Janca A. Assessment of somatoform disorders: a review of strategies and instruments. Acta Neuropsychiatr. 2003;15(4):167-79.
12. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19(4):539-49.
13. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-45.
14. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62(10):1006-12.
15. Terwee CB, Jansma EP, Riphagen, II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res. 2009;18(8):1115-23.
16. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012;21(4):651-7.
17. WHO. SCAN. Schedules for Clinical Assessment in Neuropsychiatry, version 2.1. Geneva: World Health Organization, Division of Mental Health; 1998.
18. First MB, Spitzer RL, M. G, Williams JBW. Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition. New York: Biometrics Research, New York State Psychiatric Institute; 2010.
19. Lecrubier Y, Sheehan DV, Weiller E, Amorim P, Bonora I, Harnett Sheehan K, et al. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. Eur Psychiatry. 1997;12:224-31.

20. Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. Trials. 2016;17(1):449.
21. van Tulder M, Furlan A, Bombardier C, Bouter LM. Updated method guidelines for systematic reviews in the Cochrane collaboration back review group. Spine (Phila Pa 1976). 2003;28(12):1290-9.
22. Korber S, Frieser D, Steinbrecher N, Hiller W. Classification characteristics of the Patient Health Questionnaire-15 for screening somatoform disorders in a primary care setting. J Psychosom Res. 2011;71(3):142-7.
23. Kroenke K, Spitzer RL, Williams JB. The PHQ-15: Validity of a new measure for evaluating the severity of somatic symptoms. Psychosom Med. 2002;64:258-66.
24. Interian A, Allen LA, Gara MA, Escobar JI, Diaz-Martinez AM. Somatic complaints in primary care: further examining the validity of the Patient Health Questionnaire (PHQ-15). Psychosomatics. 2006;47(5):392-8.
25. Witthoft M, Hiller W, Loch N, Jasper F. The latent structure of medically unexplained symptoms and its relation to functional somatic syndromes. Int J Behav Med. 2013;20(2):172-83.
26. van Ravesteijn H, Wittkampf K, Lucassen P, van de Lisdonk E, van den Hoogen H, van Weert H, et al. Detecting somatoform disorders in primary care with the PHQ-15. Ann Fam Med. 2009;7(3):232-8.
27. Becker S, Al Zaid K, Al Faris E. Screening for somatization and depression in Saudi Arabia: a validation study of the PHQ in primary care. Int J Psychiatry Med. 2002;32(3):271-83.
28. Muramatsu K, Miyaoka H, Kamijima K, Muramatsu Y, Yoshida M, Otsubo T, et al. The Patient Health Questionnaire, Japanese version: validity according to the MINI-International Neuropsychiatric Interview-Plus. Psychol Rep. 2007;101:952-60.
29. Corapcioglu A, Ozer GU. Adaptation of revised Brief PHQ (Brief-PHQ-r) for diagnosis of depression, panic disorder and somatoform disorder in primary healthcare settings. Int J Psychiatry Clin Pract. 2004;8(1):11-8.
30. Tebbe BB, Terluin B, Koelewijn JM. Assessing psychological health in midwifery practice: a validation study of the Four-Dimensional Symptom Questionnaire (4DSQ), a Dutch primary care instrument. Midwifery. 2013;29(6):608-15.
31. Terluin B, Smits N, Miedema B. The English version of the four-dimensional symptom questionnaire (4DSQ) measures the same as the original Dutch questionnaire: a validation study. Eur J Gen Pract. 2014;20(4):320-6.
32. Czachowski S, Terluin B, Izdebski A, Izdebski P. Evaluating the cross-cultural validity of the Polish version of the Four-Dimensional Symptom Questionnaire (4DSQ) using differential item functioning (DIF) analysis. Fam Pract. 2012;29(5):609-15.
33. Chambe J, Le Reste JY, Maisonneuve H, Sanselme AE, Oho-Mpondo J, Nabbe P, et al. Evaluating the validity of the French version of the Four-Dimensional Symptom Questionnaire with differential item functioning analysis. Fam Pract. 2015;32(4):474-9.
34. Terluin B. Factoranalyse van drie klachtenlijsten [Factor analysis of three questionnaires]. Tijdschrift voor Psychosomatische Fysiotherapie. 1999;4:25-31.
35. Terluin B. Wat meet de Vierdimensionale Klachtenlijst (4DKL) in vergelijking met enkele bekende klachtenlijsten? [What does the Four-Dimensional Symptom Questionnaire (4DSQ) measure compared to several known questionnaires?]. Tijdschrift voor Gezondheidswetenschappen. 1998;76:435-41.
36. Terluin B. De Vierdimensionale Klachtenlijst (4DKL). Een vragenlijst voor het meten van distress, depressie, angst en somatisatie. [The Four-Dimensional Symptom Questionnaire (4DSQ). A questionnaire for measuring distress, depression, anxiety and somatization]. Huisarts Wet. 1996;39(12):538-47.
37. Schmitz N, Hartkamp N, Kiuse J, Franke GH, Reister G, Tress W. The Symptom-Check-List-90-R (SCL-90-R): A German validation study. Qual Life Res. 2000;9:185-93.

38. Katerndahl DA, Amodei N, Larme AC, Palmer R. Psychometric assessment of measures of psychological symptoms, functional status, life events, and context for low income Hispanic patients in a primary care setting. Psychol Rep. 2002;91:1121-8.
39. Tyrer H, Ali L, Cooper F, Seivewright P, Bassett P, Tyrer P. The Schedule for Evaluating Persistent Symptoms (SEPS): a new method of recording medically unexplained symptoms. Int J Soc Psychiatry. 2013;59(3):281-7.
40. de Waal MW, Arnold IA, Spinhoven P, Eekhof JA, Assendelft WJ, van Hemert AM. The role of comorbidity in the detection of psychiatric disorders with checklists for mental and physical symptoms in primary care. Soc Psychiatry Psychiatr Epidemiol. 2009;44(1):78-85.
41. Christensen KS, Fink P, Toft T, Frostholm L, Ornbol E, Olesen F. A brief case-finding questionnaire for common mental disorders: the CMDQ. Fam Pract. 2005;22(4):448-57.
42. Beirens K, Fontaine JR. Development of the Ghent Multidimensional Somatic Complaints Scale. Assessment. 2010;17(1):70-80.
43. Fabiao C, Silva MC, Barbosa A, Fleming M, Rief W. Assessing medically unexplained symptoms: evaluation of a shortened version of the SOMS for use in primary care. BMC Psychiatry. 2010;10:34.
44. Budtz-Lilly A, Fink P, Ornbol E, Vestergaard M, Moth G, Christensen KS, et al. A new questionnaire to identify bodily distress in primary care: the 'BDS checklist'. J Psychosom Res. 2015;78(6):536-45.
45. APA. Diagnostic and statistical manual of mental disorders, 3d edition. Washington, DC: American Psychiatric Association; 1980.
46. Derogatis LR. SCL-90-R. Administration, Scoring and Procedures. MANUAL-II. Towson, MD: Clinical Psychometric Research; 1983.
47. Fink P, Ewald H, Jensen J, Sorensen L, Engberg M, Holm M, et al. Screening for somatization and hypochondriasis in primary care and neurological in-patients: a seven-item scale for hypochondriasis and somatization. J Psychosom Res. 1999;46(3):261-73.
48. Rief W, Hessel A, Braehler E. Somatization symptoms and hypochondriacal features in the general population. Psychosom Med. 2001;63:595-602.
49. Rief W, Hiller W. Toward empirically based criteria for the cllassification of somatoform disorders. J Psychosom Res. 1999;46(6):507-18.
50. de Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine, a practical guide. Cambridge, United Kingdom: Cambridge University Press; 2015.
51. Smits IAM, Timmerman ME, Barelds DPH, Meijer RR. The Dutch Symptom Checklist-90-Revised. Is the use of the subscales justified? Eur J Psychol Assess. 2014;31(4):263–71.
52. Sereda Y, Dembitskyi S. Validity assessment of the symptom checklist SCL-90-R and shortened versions for the general population in Ukraine. BMC Psychiatry. 2016;16:300.
53. Tomioka M, Shimura M, Hidaka M, Kubo C. The reliability and validity of a Japanese version of Symptom Checklist 90 Revised. Biopsychosoc Med. 2008;2:19.
54. Escobar JI, Rubio-Stipec MS, Canino G, Karno M. Somatic Symptom Index (SSI): a new and abridged somatization construct. J Nerv Ment Dis. 1989;177(3):140-6.
55. Rief W, Hiller W. A new approach to the assessment of the treatment effects of somatoform disorders. Psychosomatics. 2003;44(6):492-8.
56. Gierk B, Kohlmann S, Toussaint A, Wahl I, Brunahl CA, Murray AM, et al. Assessing somatic symptom burden: a psychometric comparison of the patient health questionnaire-15 (PHQ-15) and the somatic symptom scale-8 (SSS-8). J Psychosom Res. 2015;78(4):352-5.
57. Hart DL, Werneke MW, George SZ, Deutscher D. Single-item screens identified patients with elevated levels of depressive and somatization symptoms in outpatient physical therapy. Qual Life Res. 2012;21(2):257-68.